

DOCUMENT RESUME

ED 177 461

CG 013 914

AUTHOR Rezmovic, Eva Lantos; And Others
 TITLE In Quest of Meaningful Evaluation Results: Lessons From a True Experiment.
 PUB DATE 3 May 79
 NOTE 24p.; Paper presented at the Annual Meeting of the Midwestern Psychological Association (51st, Chicago, Illinois, May 3-5, 1979)
 EDRS PRICE MF01/PC01 Plus Postage.
 DESCRIPTORS *Agencies; *Correctional Rehabilitation; *Data Collection; Employment Programs; *Evaluation Methods; *Field Interviews; Information Seeking; Rehabilitation Counseling; *Research Problems; Social Problems
 IDENTIFIERS *Ex Offenders

ABSTRACT

The emphasis on the mechanics of research techniques has often overshadowed other aspects of evaluation research that are determiners of research quality. Many problems encountered by program evaluators are not solvable by the information provided in research texts. Some of the practical problems encountered and dealt with during an experimental evaluation of a criminal justice program are described. Issues revolving around budgeting, randomization, data collectors, interview payments and evaluation of black box treatments are addressed. (Author/BMW)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED177461

In Quest of Meaningful Evaluation Results:
Lessons from a True Experiment

Eva Lantos Rezmovic
Northwestern University

Thomas J. Cook
Research Triangle Institute

L. Douglas Dobson
Northern Illinois University

CG013914

Paper presented at the Midwestern Psychological Association, Chicago,
Illinois, May 3, 1979.

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Eva L. Rezmovic

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Introduction

In judging the quality of social program evaluations, critics tend to focus on the technical features of research conduct. For example, large sample size, implementation of random assignment, reliable and valid outcome measures, and appropriate statistical analyses are commonly used indicators of "good/rigorous" research methodology. A study that meets these criteria is likely to receive professional recognition as a prototype of good research practice.

While technical competence is indeed important, the emphasis on the "mechanics of research technique" (Roth, 1970, p. 556) has often overshadowed other aspects of evaluation research that are equally important determiners of research quality. Many problems encountered by program evaluators are not solvable by following the dicta of research texts; nor are their effects estimable through statistical manipulations. Nevertheless, the way most studies are written up, one can easily get the impression that aside from the technical difficulties of designing and implementing the evaluation, it was smooth sailing throughout. This impression is clearly false, particularly in the case of applied policy research.

In this paper we will try to communicate the experience gained from an experimental evaluation of a criminal justice program. We will not dwell on the technical details of the study as this has been reported elsewhere, (Cook, Dobson, & Rezmovic, 1978). Rather, we will indicate some of the practical problems encountered during the study, how we dealt with them, and the lessons learned. While the experiences are drawn from a particular study, it is hoped that the discussion will be useful to others involved in similar research.

The Experiment

The setting for the evaluation is Chicago's Safer Foundation, a private corporation whose expressed goal is to help ex-offenders help themselves. It seeks to accomplish this by focusing on the employment needs, as well as the social/personal needs, of adult ex-offenders who are released from federal, state, and county prison systems into the Chicago metropolitan area.

Our 2-year, LEAA-funded evaluation began in February, 1977 and was designed to assess the relative effectiveness of two service programs operated by the foundation. One program, Operation DARE, deals only with the employment needs of ex-offenders: it provides job counseling and job referral services to over 2000 ex-offenders a year. The other program, known as CHALLENGE, focuses on the social reintegration of ex-offenders into community life, and it is manned primarily by volunteer workers. The volunteers work with approximately 1000 parolees each year in one of three treatment models. These treatment models, which constitute the three experimental conditions of the evaluation, are the following: (1) Staff Associate model: staff associates are private citizens who volunteer to be matched with an ex-offender for a period of one year. During that year, the staff associate serves as a resource person, ombudsman, and listening ear to the ex-offender client. Together, they work on the ex-offender's immediate needs (e.g., housing, clothing, job) and short- and long-term goals (e.g., schooling); (2) VISTA model: VISTA counselors are volunteers of the national ACTION program who have been placed with the Safer Foundation. These counselors are themselves ex-offenders, functioning under similar economic straightens and cultural values as their clients. As examples of

ex-offenders who have "made it," they are thought to be in an excellent position to offer guidance to recently released parolees. In contrast to the yearly, one-to-one relationships of the Staff Associate model, VISTA counselors work with caseloads of 15-20 clients for three month intervals. The counseling functions of the VISTA's, however, are similar to those of Staff Associates--they serve as crisis intervenors, practical problem solvers, representatives, and sounding boards to their clients; (3) Paraprofessional model: Paraprofessionals are employed directly by the Safer Foundation and are assigned to work with clients for a six-month period. Each paraprofessional carries a caseload of 25-30 clients and provides services similar to those of the other models. In general, the paraprofessional counselors are also ex-offenders.

Between May and September, 1977, 367 new ex-offender clients of the Safer Foundation were randomly assigned to the DARE program or to one of the three treatment components of the CHALLENGE program. It should be noted, however, that the counseling and support services offered by CHALLENGE are supplementary to, not independent of, the employment services provided by DARE. Hence, the purpose of the evaluation is to determine the incremental/marginal gains associated with participation in one of the three treatment models of the CHALLENGE program.

At this point, we are in the process of analyzing the data from the experiment. The discussion to follow will, therefore, not be concerned with the results of the analyses but rather discuss some of the issues and problems we confronted during the implementation of the experiment.

At the Beginning: Setting the Budget -- One of the key decisions made at the outset of the project was to incorporate the research budget into

the overall operating budget of the CHALLENGE Program. Several reasons prompted this decision. Primarily, we believed that combining the research and operational budgets would serve to strengthen the commitment of the agency to a successful completion of the experiment. They would, in a sense, be a "partner" in the experiment and have a vested interest in seeing it through to completion. We also thought that the single budget would promote the utilization of the evaluation results by the agency: since the agency was spending its own money, it would be more likely to use the results. Finally, we felt that the single budget would be more palatable to funding sources since part of the proposal for the experiment emphasized the use of the project as a vehicle for developing an "in-house" evaluation capability within the agency. As it turned out, what seemed like a good idea in theory did not work out smoothly in practice.

The major adverse consequence of the decision was to place the research project in direct competition with program operations for scarce resources. In discussions with agency management, the issue of program needs vs. research needs invariably cropped up. The issue became especially salient toward the end of funding periods, when the agency had not yet received confirmation of funding and faced meeting a payroll. At these junctures, agency management made it clear that program survival outweighed the need to preserve the research budget. As one person put it, "without the program there's nothing to evaluate."

In retrospect, it seems obvious that the agency environment was ripe for producing tensions over budget allocations. Small, non-profit organizations like the Safer Foundation rarely possess a solid base of continuing financial support. They spend a tremendous amount of time and effort preparing appeals for funds; they no sooner finish one funding application when

they are gearing up for another. The pressure to get funds is always there. Under these conditions, available funds, research or otherwise, became "fair game" in the struggle to keep the program going.

Fortunately for us, the necessary funds to continue program operations eventually came through in time. We have no doubt, however, that had they not come through, the evaluation's budget would have been cut. Without malice, the agency would simply have used any available funds as a means of survival.

It is too early to judge the merits of the single budget idea vis-a-vis other considerations, such as promoting the utilization of evaluation results. Nevertheless, our experience to date has caused us to seriously question the wisdom of the approach as a budgeting strategy. Separate budgets not only permit more accurate accounting, they also circumvent misunderstandings about budget priorities and reduce uncertainties about the availability of research funds throughout the period of the study. Completing an experimental evaluation is difficult enough without the added complications and tensions entailed in periodic budget negotiations. Hence, our experience suggests that it is better to, before the fact, establish formal, written agreements specifying what funds will be spent exclusively for research and, in as much detail as possible, how these funds will be spent.

Collecting the Data: Problems in Hired Hand Research

One problem that evaluation reports faithfully ignore concerns the problems of conducting "hired hand research". As considered by Roth (1970) hired hands are individuals who function mainly as data collectors or processors in an evaluation study, and who have little or no part in

developing the evaluation plan or in reaping benefits from carrying out a high quality study. Such individuals have so little at stake in the research that it becomes easy for them to take shortcuts in completing tedious tasks, to sacrifice accuracy for speed when they are under time pressure, and to blatantly cheat or lie at the expense of valid data when it is expedient for them to do so. According to Roth, when hired hands realize that they are there only to do the researcher's "dirty work," then "some form of cheating, carelessness, distortion, or cutting of corners in the collection or processing of research data...is not abnormal or exceptional, but rather is exactly the kind of behavior we should expect from people with their position in a production unit" (p. 545). Having conducted a study which relied on hired assistants, we can draw upon the experience to make some comments about the reasonableness of Roth's speculations.

Regrettably, there do appear to be grave data corruption risks when conducting hired hand research. When we discovered midway through the project that both interviewers had falsified interviews which purportedly took place either in the field or at the agency during weekends and evenings, we were taken aback. A single client's denial to having been interviewed before, when a completed interview form had been turned in on him, was sufficient to precipitate concern that the project had been sabotaged. While a random call-back procedure had been in operation throughout the study to verify that interviews which were not held at the agency during weekday office hours did indeed take place, we were not often successful in contacting clients. As we repeatedly found out, many clients either spent little time at home, had no phone, or had no permanent place of residence. It was because of these difficulties that we instituted field,

evening, and weekend interviewing to begin with: clients were so difficult to locate by telephone that interviewers were sent to their homes to either schedule an agency interview (for which the client received \$10) or to conduct an interview on the spot (for which the client received \$5).

During the weeks and months following our discovery of the forged interview, we managed to contact most of the clients whose interviews were suspect. Through tenacious telephone calling and by asking each client who was interviewed thereafter whether they recalled the prior interviews, we deduced that only a handful of interviews were falsified. Also, during lengthy interviews with the culprits we had them identify, as best they could from recall, the counterfeit interviews. One of the offenders even had the temerity to request payment for his cooperation!

While the forging of interviews did not seem to have been rampant, it does give one pause about the reasons why data corruption occurred at all. In attempting to understand the conditions that facilitated its occurrence, the following factors can be identified from the CHALLENGE study:

(1) Frustration level of the task: As noted above, the task of contacting clients and arranging interviews was not an easy one. Ex-offenders tend to be highly mobile during the period following prison release, and relatives or friends don't readily volunteer information to strangers on their whereabouts. Occasionally, a tracker/interviewer would spend days trying to establish a client's whereabouts (via Safer program staff, the client's parole officer, County Jail, telephone calls, home visits, etc.) without making a single direct contact; and consequently, without producing a single interview. Although the tracker/interviewers were repeatedly assured that their job performance was

evaluated on the basis of the effort they exerted and not on the number of interviews completed, it is not difficult to understand how these assistants felt. If the completion of the interview may be viewed as the internal reinforcement of the job, then the reinforcement schedule was very meager at times. Hence, the trackers/interviewers took it upon themselves to reinforce their strenuous, though futile, efforts by fabricating an occasional interview.

(2) Personal problems of hired hands: It is generally agreed that valid research requires a committed research team (Fairweather & Tornatsky, 1977). However, commitment is neither a dichotomous variable nor one that stands alone as the determiner of job performance. While a high level of commitment to one's job and to the integrity of the study may be necessary preconditions for valid research, they are not sufficient. Factors external to the research study and to the hired hands' attitudes toward it may compete with commitment in influencing job performance. Specifically, these external factors may be in the form of personal problems.

In the Safer experiment, the hired assistants were both "committed" to the research effort and beset by personal problems which interfered with their job functions. These problems included marital discord, financial pressures, extensive family obligations, and alcohol usage. The effect of these problems was that the hired hands would sometimes take care of "personal business" during their trips into the field rather than perform their tracking duties. To make it appear that their time was productively spent, they then forged some interviews. Despite these actions, these assistants also maintained that their jobs were important to them, that the evaluation and its goals were of interest to them, and that they had been motivated to do well on the job.

When conducting research on ex-offenders using other ex-offenders as part of the research team, there are many unknowns. It is thought that by hiring individuals who are indigenous to the community of interest, it will be easier to maintain contact with research clients, to establish rapport with them, and to obtain valid data from them (Argyris, 1971; Riecken & Boruch, 1974). On the other hand, ex-offenders (and others of low economic status) often have more social and personal problems than the general population. Depending on the nature and quantity of such problems, they may prevent hired hands from performing with integrity in their jobs. This appeared to be the case in the Safer experiment. Whether a more general case can be made regarding the relationship between the personal problems of hired hands and corruption of research data depends on the candor of other evaluators in reporting their experiences.

(3) Concrete benefits to cheating: When failure to perform the job can result in a windfall for the hired hand, the motive to cheat becomes intensified. Since all respondents in the Safer study were to have received payments for their interviews, all cases of falsified interviews resulted in the hired hands' pocketing the money. Although the total amount profited in this way was minimal (less than \$100), the hired hands did admit that their financial obligations were so great that any amount helped.

Depending upon the setting and target population of a study, the problem of financial abuse being coupled with data corruption can be handled with varying degrees of control. Among the safeguards that were built into the Safer research to prevent this problem from occurring were the following: Clients who were interviewed at the agency during the week were paid with checks. To enable these clients, very

few of whom had bank accounts, to cash these checks without difficulty, special arrangements were made with a bank located across the street from Safer. Because of these provisions, there was no incidence of falsified data for weekday agency interviews. Field, evening, and weekend interviews posed a different sort of problem, however. Because in these cases clients would have been unable to cash the interview check immediately, there was concern that they would be reluctant to grant the interview. Therefore, these clients were paid in cash, and signed receipts were obtained from them. While the idea of signed receipts seemed to be a good one at the time, it was clearly not foolproof: In reviewing the receipts collected, we failed to note the cases where there were handwriting discrepancies from one signature to the next across different waves of interviews; the sham was exposed in our later investigations.

One might argue that the obvious preventive would have been to screen carefully in recruiting our hired hands. We thought so too--much to our chagrin. The main perpetrator was a Chicago native with a bachelor's degree, coursework completed on a master's degree in criminal justice, and assuring recommendations from former professors. The person also had experience in field interviewing and, as an ex-convict, was ostensibly committed to the project, "as a way of helping other ex-convicts make it outside prison". A "perfect" candidate for the job--this person not only instigated the sham, he "trained" his successor!

At this point, we cannot offer a definitive prescription on how to avoid the hazards involved in using hired hands. In this experiment, we felt constrained to use people indigenous to a particular locale; we have attempted to describe aftermath of that choice. Hopefully, our experience will serve to caution others planning similar resources and aid the search

for potent methods to neutralize data corruption threats. The alarming, yet very real problem of how social structures lead to data corruption in an area that has been grossly neglected in the evaluation literature.

Compensating Research Subjects

It is not difficult to think of reasons why study participants should be paid for providing research data. Foremost among these is that payments should reduce attrition, increase response accuracy, and impel respondents to take the research more seriously. Furthermore, reimbursing clients for their time and effort is an ethical gesture (Weiss, 1975).

Aside from the burden of increased administrative costs, there appear to be few reasons for not paying subjects. However, one possible validity threat is that subjects will be more likely to respond to the demand characteristics of the situation. That is "(i)f a person is being paid, he may feel constrained to give value in the form of obligingly invalid data" (Weiss, 1975, p. 375).

While the weight of opinion may be on the side of offering compensation for research participation, empirical evidence on the effects of monetary incentives is meager. Since the Safer experiment was not designed to test hypotheses about paying clients for research interviews, we cannot make any definitive statements about its impact. However, some of our observations on the subject demonstrate its complex nature and provide case material for the evaluator who may be inclined to decide that, at worst, payments for research participation are innocuous.

Our first indication that interview payments may not have totally beneficent value occurred very early in the project. Intake interviewing

had been in operation for approximately four weeks and appeared to be proceeding at a satisfactory pace. Then, the head of the Staff Associate matching unit approached us to request that research interviews be postponed until clients had been matched with citizen volunteers. The reason: Once clients received their \$10 interview payments, they were reluctant to return to the agency for the matching session. Whether or not this contention was correct, and if it was, why, was it not directly ascertainable. However, because of the dire poverty of many newly released prisoners, it was conceivable that some clients volunteered for treatment because they sought the interview money, and disregarded treatment once the money was in hand. Others may have volunteered in good faith, but placed the program secondary when some of their immediate survival needs were met by the DARE program and the interview payment. While there was no hard evidence to support either of these possibilities, Safer staff strongly felt that our interview payments posed as a disincentive to clients to return for matching. Not wishing to jeopardize either good relations with the program or the power of the evaluation by risking the possibility that we were contributing to attrition from treatment, we struck a compromise: Intake interviews were delayed one week to provide additional time for the program to consummate matches between clients and citizen volunteers. Any lengthier delay, it was felt, would run the risk of increasing attrition from the research.

It is commonly thought that if monetary incentives do act as a deterrent to attrition, then they should be most effective when research participants are poor. However, this belief fails to acknowledge that there are two types of attrition that evaluators must be concerned with: Attrition from treatment and attrition from the research. Furthermore, the Safer experience raises the possibility that compensation may differentially

affect these two types of attrition. If paying research subjects could simultaneously reduce the research drop-out rate and contribute to the treatment drop-out rate, then its role is self-defeating. At the very least, there should be recognition that payments may actually exacerbate attrition problems.

A second concern regarding interview payments dealt with its effects on the study's external validity. If, in fact, the decision of some ex-offenders to participate in the treatment program was mediated by the prospect of receiving \$10 for each of three interviews, then the study sample may not have been truly representative of the target population. In hopes of avoiding this problem, the research and program staffs had worked out in advance a procedure for recruiting clients into the treatment and evaluation programs. This procedure required that only after clients were informed of the treatment alternatives and had volunteered to participate in them would they be informed of the research project. In this way, the offer of \$10 interview payments would not serve as bait for entering treatment. Unfortunately, there are two reasons to doubt the efficacy of this precaution: (a) Program staff did not always adhere to the prescribed delivery routine; particularly during the early days of the evaluation period, when the procedure was still new to them; (b) Word about the research project got around among ex-offenders so that some clients knew about the availability of an "easy \$10" even before coming to the agency. In fact, on two separate occasions following the termination of random assignment, ex-offenders came to the research office saying that they heard we were giving \$10 handouts and wanting their share of the boon.

A third area in which the benefits of interview payments were questioned concerned the extent to which they impelled subjects to take the research

more seriously. While the absence of comparison renders the observation merely speculative, it is doubtful that interview payments had a significant effect on the attitudes of our ex-offender subjects toward the research project. For example, our experience was that many clients who had found jobs no longer had plenty of free time to come in for interviews. Since their time was scarce and the \$10 interview payments were no longer of such great value as when they were unemployed, much persuasion was needed to convince such clients that the interviews were important enough for them to set aside time for. At the other end of the spectrum were clients who remained unemployed and appeared interested in the research interviews only because of the compensation they would receive. We have a number of recorded instances of clients either requesting an "advance" on their next interview payment, or inquiring about the possibility of scheduling early interviews. In both cases, need for immediate cash, and not seriousness about the research, was the motivating force behind clients' interest in the interviews.

Fourth, despite careful explanation that the research project was independent of the programs at Safer, a number of ex-offenders did not differentiate between the two. When asked about specific ways in which the DARE and CHALLENGE programs had been helpful to them, these clients cited the \$10 interview payments. Whenever an evaluation office is located at the same site as the program it is evaluating, there is a risk that subjects' opinions will be affected by their experiences with the evaluation. Therefore, in gathering attitudinal/perceptual data from research subjects, on-site evaluators should be aware that they may be evaluating their own effects on clients (e.g., via interview payments, as in the case of the Safer study) in addition to those of the program.

Finally, we had other incidents that perhaps were predestined given the nature of the study and its locale. One interviewer was mugged (or so he claimed) while in the field tracking clients and lost twenty dollars in interview payment money. The research project borders on an area of Chicago known as skid row. Somehow, word of the interview payments reached some of the "residents" of this area. We had several drunks come to the agency, claim they were ex-offenders, fill out the initial screening forms, sit through the CHALLENGE orientation, and stagger into the research office to be interviewed. Minor damage was done (two interviews completed) before we were tipped off to the sham and could tighten the screening process whereby people were admitted into the CHALLENGE orientation.

To date, the effects of monetary compensation on attrition from treatment and research, external validity, attitudes of research participants, and response bias are unknown. While intuition concerning the benefits of interview payments may be borne out by research findings, such findings have yet to be produced. We raise these reservations about the value of interview payments because the issue is not as clearcut as it might first appear.

Black box intervention

Through painful experience, social researchers have come to recognize the importance of describing, monitoring, and measuring treatment delivery. The experience has been painful because we have too often found ourselves at a loss to explain why a particular intervention did or did not produce the expected effects. In response to prior neglect of documenting and measuring what took place during the treatment period, writers on evaluation methodology are becoming increasingly vocal and adamant in calling for

studies which focus on program components as well as outcomes. At the same time, they have written little about the practical problems of obtaining process data in impact evaluations.

The task of collecting process data is most easily dealt with when the evaluator is also the treatment administrator and thus has maximum control over experimental conditions (Fairweather & Tornatzky, 1977). This task becomes more difficult when the roles of evaluator and treatment administrator are handled by different individuals. Nevertheless, very satisfactory attempts at assaying treatment strength may be made if treatment delivery takes place in a closed setting (e.g., prison, hospital, school). The least tractable situation for monitoring treatment--the one which presented itself in the Safer experiment--arises when the intervention is of a human service nature and is delivered in the community by nonresearch personnel. In the VISTA and Paraprofessional conditions of the Safer study, counseling staff from the agency met with ex-offender clients in the clients' homes; in the third condition, Staff Associates met with clients on "neutral ground" (e.g., parks, restaurants, museums) in the community.

The problems for monitoring strength and integrity of treatment were manifold. In particular, we were faced with the following limitations: (a) If meetings took place in conformance with the treatment plan, there should have been, overall, more than 1500 client-counselor meetings during the 2-year evaluation period. Moreover, these meetings were held on a one-to-one basis in diverse community settings. Given this decentralized form of treatment delivery, considerable staff and resources were necessary to monitor even a fraction of these meetings. Such resources were not available in the Safer study, and are undoubtedly just as scarce in other evaluations; (b) In point of fact, client-counselor meetings numbered far fewer than 1500. The reasons for this varied, but failure to meet was

often due to counselors' inability to contact their clients. In the case of VISTA's and Paraprofessionals, unsuccessful attempts to reach the client by telephone would generally result in a home visit. Such unarranged visits, however, did not often find the client at home. Had evaluation staff accompanied counselors on these trips, much staff time would have been expended on unproductive efforts to simply make contact with clients;

(c) Most counselors hired by the Safer Foundation are themselves ex-offenders, native Chicagoans, and fairly unhesitant about visiting poor, black, high-crime neighborhoods in Chicago. Because most newly released prisoners reside in such neighborhoods, this is where meetings between clients and their VISTA and Paraprofessional counselors took place. By contrast, the evaluation staff was not as "street wise" as the agency counselors and would not have easily fit into their milieu. This served as a further obstacle to treatment monitoring. While a "street-wise evaluator" would be the ideal resource person for studies involving community treatment of offender samples, this type of individual remains a relative rarity;

(d) Because counseling sessions were held individually in private homes, restaurant booths, etc., it would have been awkward and distracting for evaluation staff to be present. In fact, VISTA and Paraprofessional counselors were not permitted to fill out the necessary agency forms--or even to take notes--during counseling sessions because CHALLENGE administration felt that this would interfere with the rapport-building purpose of the meetings. It was made clear to the evaluation staff that our presence in meetings with clients would be threatening to both clients and counselors and detrimental to maintaining a trusting relationship between them. In effect, we were not welcome to observe the contents of the black box; (e) We were also not able to listen to the contents of the

black box. The evaluation project's finances did not permit purchase of recorders to tape what was discussed during counseling sessions. With over 70 counselors meeting periodically with clients in the field, buying enough records to monitor treatment meaningfully would have been prohibitive. Besides, counselors were not receptive to this idea, either, claiming that recorders would bias the nature of their interactions with clients.

Combined, the factors cited above formed a potent obstacle to assessing the components of treatment in the CHALLENGE program. While we did interview clients to inquire how many times they met with their counselors, what kinds of topics were discussed, whether and how the counselors were helpful to them, and whether they felt that the counselors cared very much about them, such questions provided only crude information on treatment delivery. Responses were subject to the distorting effects of recall errors, which is a drawback in any type of retrospective reporting. Additionally, our ability to uncover the various dimensions of treatment through specific description of the counseling sessions was curtailed by clients' inarticulateness. Hence, only general information was obtained from these indirect attempts to appraise treatment quality/intensity.

The question that arises is what the evaluator should do when confronted by such impediments to unlocking the black box of treatment implementation. According to some, such black box interventions should not be implemented in the first place (Rossi, 1978), if implemented they should not be evaluated (Rossi, 1978), and if evaluated and found to be ineffective, their results should be ignored (Sechrest, 1979). While these suggestions may have some merit, there should be available some more constructive advice on how to handle problems of monitoring and measuring treatment imposition. In the case of the Safer study, for example, the evaluation was conducted with full

awareness of the importance of specifying, monitoring, and measuring treatment. However, the chasm was great between our knowledge of scientific desiderata for conducting such a powerful evaluations and our ability to apply it in the field setting. The dilemma was aptly summarized by Mann (1971): "unfortunately, the greater the scientific precision in demonstrating the effectiveness of a given method, the greater the likelihood that the experiment bears little resemblance to the method as it is normally employed by practitioners. On the other hand, when a procedure is tested as it is used by practitioners, it is almost impossible to describe the change process in a precise and scientific manner." (p. 178).

Our experience exemplifies the problem of monitoring and measuring treatment implementation, but does not resolve it. This problem, we feel, is an intractable one to many applied researchers. It is time that the evaluation research community responded to the need for developing practical guidelines on how to tap the contents of treatments.

Conclusion

We have drawn on our experiences in conducting a true field experiment in order to direct attention toward a facet of evaluation not normally discussed in standard research texts or in evaluation reports. In particular, we have tried to raise some of the practical, "real world" types of issues that can arise in the course of managing an evaluation project. These issues are not unique to controlled studies. Indeed, a major point of this paper is that a strong experimental design cannot stand alone: the successful implementation of randomization (and the Safer study did effect a successful random assignment strategy) is but one step toward obtaining valid and meaningful evaluation results.

We also wished to highlight the need for communicating not only research methodology, but also research experience. The faithful recording of research experiences would provide a rich lore for the evaluation community. Researchers would be forewarned about the costs and benefits associated with certain strategies, such as paying interviewees. Likewise, they could capitalize on the lessons learned in dealing with various research management issues. The problems connected with issuing hired hands is an obvious example; the journals could be mined for safeguards against data corruption threats. Perhaps the main value of such annals would be to discourage the untimely initiation of evaluation studies. Researchers could survey prior experience as a means of assaying the chances of successfully completing an evaluation. The survey may prompt a decision either to postpone or terminate a prospective study; either decision might avoid wasting resources in a foredoomed venture.

Too little has been written about the kinds of procedural problems evaluators encounter in attempting to maintain the integrity of experiments. Many of these problems, however, are common enough to merit public airing. We look forward to a change in conventional reporting practices and hope that this paper has contributed toward that end.

References

- Argyris, C. Creating effective research relationships in organizations. In F. G. Caro (Ed.), Readings in evaluation research. New York: Russell Sage, 1971.
- Chromy, J. R., and Horvitz, D. G. The use of monetary incentives in National Assessment household surveys. 1974 Social Statistics Section: Proceedings of the American Statistical Association. Washington, D.C.: American Statistical Association, 1975, 1/1-179.
- Cook, T. J., Dobson, L. D., and Rezmovic, E. L. Experimental evaluation of the CHALLENGE program: First year report. Chicago: Safer Foundation, 1978.
- Fairweather, G., and Tornatzky, L. Experimental methods for social policy research. New York: Pergamon, 1977.
- Hubbard, R. L., Rachal, J. V., and Ginzburg, H. M. A preliminary examination of the effects of compensation for interviews on the response quality in a study of drug treatment programs. Research Triangle Institute, unpublished.
- Lansing, J. B., Ginsburg, G. P., and Braaten, K. An investigation of response error. Urbana, Ill.: University of Illinois Press, 1961.
- Mann, J. Technical and social difficulties in the conduct of evaluative research. In F. G. Caro (Ed.), Readings in evaluation research. New York: Russell Sage, 1971.
- Riecken, H. W., and Boruch, R. F. (Eds.). Social experimentation: A method for planning and evaluating social intervention. New York: Academic Press, 1974.
- Rossi, P. H. Measurement and decision making. Paper presented at The 1978 Educational Testing Service Invitation Conference, October, 1978.
- Roth, J. Hired hand research. In N. K. Denzin (Ed.), Sociological methods: A sourcebook. Chicago: Aldine, 1970.
- Sechrest, L. Personal communication, January, 1979.
- Weiss, C. H. Interviewing in evaluation research. In E. L. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. 1. Beverly Hills, Calif.: Sage, 1975.

Footnotes

The authors wish to thank the Safer Foundation for their assistance in the design and implementation of the evaluation described in this paper.